AD_____

Award Number: W81XWH-09-1-0694

TITLE: Assessing the role of copy number variants in prostate cancer risk and progression using a novel genome-wide screening method

PRINCIPAL INVESTIGATOR: Donna Lehman, PhD

CONTRACTING ORGANIZATION:
University of Texas Health Science Center
San Antonio TX 78229-3901

REPORT DATE: October 2010

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
01-10-2010	Annual	15 Sep 2009 - 14 Sep 2010

4. TITLE AND SUBTITLE

Assessing the role of copy number variants in prostate cancer risk and progression using a novel genome-wide screening method

5a. CONTRACT NUMBER

5b. GRANT NUMBER

W81XWH-09-1-0694

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Donna Lehman
Robin Leach
August Blackburn

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

University of Texas Health Science Center

San Antonio TX 78229-3901

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Individual copy number variations in the genome may play a substantial role in influencing trait variation, yet due to technical limitations they have been understudied, and little is known about this new class of variant, including their distribution in most human populations and impact on common diseases. The goal of the current research is to screen the autosomal genome for these variants in constitutional DNA to assess their role in risk of development of prostate cancer and then evaluate any direct effect on the prostate. Using an array-based method we have identified heritable copy number variable regions that significantly differ between prostate cancer cases and controls of Mexican American origin. We have developed and conducted independent assays to analyze and validate CNVs identified using the array-based method and have validated 8 of 8 attempted. We have determined that a gene set comprised of genes in these CNV regions is statistically enriched for genes shown to be differentially expressed in prostate tumor and normal tissue. This supports our hypothesis that heritable structural variation may affect risk for PCa and/or its progression.

15. SUBJECT TERMS

Heritable copy number variation, prostate cancer, gene set enrichment

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			USAMRMC
U	U	U	UU	9	19b. TELEPHONE NUMBER *(Include area code)*

Table of Contents

INTRODUCTION

Prostate cancer is known to have a strong genetic component. Thus, the identification of the heritable genetic alteration(s) that precedes or increases susceptibility to somatic cancerous changes in the prostate could likely lead to improved identification of high risk individuals for early screening and possibly to new treatment strategies. Standard methodologies, including linkage analysis in familial prostate cancer patients and genome-wide single nucleotide polymorphism (SNP) screening have not identified sufficient genetic alterations to account for the hereditary component of prostate cancer. Recently, it has become apparent that structural variation comprises the majority of the diversity of human genomes, much more than SNPs, and may play a significant role in disease susceptibility and resistance. Since CNV regions often contain genes, parts of genes, or regulatory regions, they could result in different levels of gene expression. In addition, through deletion between genes or insertion of duplicated sequences into a gene, CNVs may also contribute to creation of new genes. Thus, they may play a substantial role in influencing trait variation, yet due to technical limitations they have been understudied, and little is known about this new class of variant, including their distribution in most human populations and impact on common diseases. The goal of the current research is to screen the entire autosomal genome for these variants in constitutional DNA to assess their role in risk of development of prostate cancer and then evaluate any direct effect on the prostate.

BODY

We previously screened the entire genome of 100 Hispanic prostate cancer subjects and 67 Hispanic controls for copy number variants using an Infinium-based array by Illumina that covered all published CNVs as well as an additional ~13,000 regions not previously covered on SNP arrays. These regions include segmental duplications, megasatellites, and regions lacking SNPs. Coverage with this tool was provided by 44,220 SNPs or non-polymorphic probes representing ~29,000 segments, 15,559 of which are non-redundant segments. In collaboration with DeCode Genetics, the microarray genotyping data underwent preprocessing to remove noise and artifacts using DeCode´s unique protocol based on in-house data models and analytical methods developed using a large body of proprietary data for the CNV chips. Next, the number of copies of each “allele” was estimated using information from the intensity values for each probe. The association between prostate cancer and each polymorphic marker was tested using logistic regression analysis (using a logit link function). A likelihood ratio test was performed comparing the null hypothesis of no association to the two-sided alternative hypothesis of association. In order to minimize the effects of confounding, relevant covariates such as age were included in the model. The model was also adjusted for potential confounding by admixture (genetic population substructure) using principal components methodology. We used an additive genetic model to model the effect of the CNV, such that each additional copy of a variant would increase (or decrease) the trait by the same amount and used a Bonferroni correction in interpreting statistical significance. Using a conservative Bonferroni corrected significance threshold of $p \leq 10^{-6}$, which correlates with an experiment-wise p-value of 0.05, we observed 13 unique probes to be significantly associated with prostate cancer. We observed 25 associated CNV loci at a significance threshold of $p \leq 10^{-5}$.

In our Statement of Work, we proposed to evaluate the top 25 associated loci in a larger dataset of cases and controls using quantitative PCR over the first project period. In the process of designing and developing assays for the 25 CNVs we have learned that the DNA sequence in a number of these regions defy successful assay design as they are complex and contain non-unique sequence. We have designed assays and completed this evaluation for 8 of

the loci in 204 cases and 437 controls. For 2 of the CNVs tested (PTPRK and HINT1) the qPCR assay did not directly overlap the region of interest. We tested for association using logistic regression with age and the counts at the CNV (i.e., additive genetic model) included as covariates. The results of analyses are shown below in table 1.

Table 1. Association of CNVs in SABOR Mexican-American subjects

CNV	chromosome	OR	OR 95% CI Lower	OR 95% CI Upper	LR. P-value
AUTS2 hs04946628	8	3.44	1.09	10.84	0.021
PTPRK hs06149836	6	0.89	0.27	3.00	0.859
HINT1 Hs03548981	5	0.51	0.12	2.15	0.344
KIAA0125 Hs03069453	14	0.40	0.16	1.01	0.046
ADAM6 Hs07100777	14	0.28	0.12	0.68	0.002
PPEF2 Hs01217789	4	9.24E-07	0	Inf	0.067*
PTPRN2 Hs04337897	7	0.836091	0.151562	4.612307	0.836
UGT8 Hs04833901	4	1.009264	0.680432	1.49701	0.963

*Note: CI includes infinite (no variants observed in cases); Fisher's exact p-value=0.264

As shown in table 1, we observed nominal association of 3 CNVs with prostate cancer. The CNVs located on chromosome 14q near the KIAA0125 and ADAM6 genes are in an immunoglobulin region which is a site of rearrangement in blood cells. Therefore, we interpret these CNVs to be somatic changes as opposed to heritable, germ-line CNVs. This region is known to be copy number variable in tumors and has recently been shown to be differentially variable between prostate tumors of African American subjects and those of Caucasian subjects, leading the authors to comment that there may be "molecular alterations at the level of gene expression and DNA copy number that are specific to African American and Caucasian prostate cancer and may be related to underlying differences in immune response"[1]. We are currently examining this region in the Caucasian subjects of the SABOR for comparison. A CNV located upstream of AUTS2 gene was associated with PCa in the larger dataset. There is no known biological connection between AUTS2 and PCa. However, this CNV is situated within a segmental duplication with the nearly identical region located within the SLCO5A1 gene. SLCO5A1 is an organic anion transporter shown to have increased expression in tumors and metastases from prostate cancer[2]. This example of a segmental duplication demonstrates the complexity of CNV analyses which we have discovered over the first year of this project. We plan to examine expression of these 2 genes in prostate normal and tumor tissue.

Our results to date have shown that some of the CNVs we previously identified are indeed rare. The CNV located just upstream of HINT1 was observed in only 1.8% of the subjects tested. HINT1 is a known tumor suppressor gene and therefore an increase in copy number (if it correlates with increase in expression) would conceivably decrease risk for cancer.

The observed odds ratio is consistent with this hypothesis; however, our results were not statistically significant. This could possibly be due to the low number of subjects with the variant. In addition, the effect of the variant allele counts may not be additive with a copy number of 1 not having adverse effects. Conversely, the qPCR assay may not have the same specificity as the array given that the assay did not directly overlap the same region. If breakpoints differ between subjects bearing the CNV, each assay will differ in copy calls. The CNV located in the PPEF2 gene was also rare with 4 control subjects and zero cases bearing the copy number increase. PPEF2 has recently been shown to be a negative regulator of apoptosis signal regulating kinase-1 (ASK1) and its expression was correlated with growth, proliferation, or neoplastic transformation, making it a candidate for cancer predisposition[3]. It is also differentially expressed in prostate tumor and normal tissue as shown below. Given that measurable PCa risk may be due to the burden of multiple genetic risk factors, each with small effect, testing for overall CNV burden may be necessary to detect the effects of rare CNVs[4].

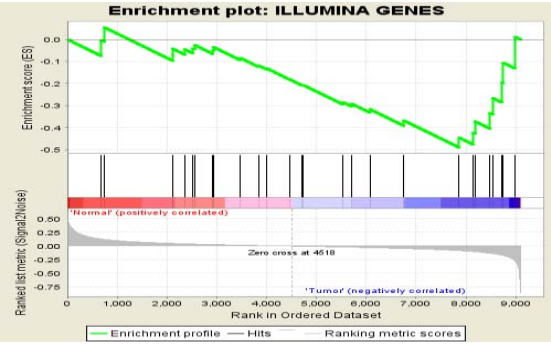
As mentioned above, a number of the associated CNV regions could not be validated using the method of quantitative PCR. We sought to find a means to help predict which gene or genes in those CNV regions may be affected by the variant and influence prostate cancer risk, thereby prioritizing genes for further analysis. Task 2 of our SOW is to identify candidate genes from CNV regions and assess any effects on expression of these genes. For this task we proposed to select candidate genes by comparison of current genomic database information with the current literature. We have decided to augment this approach by conducting a gene set enrichment analysis (GSEA) using published expression data from prostate tissues. Briefly, the closest genes to the 25 most significantly associated variants were identified as a gene set. Next, we used Robust Probe-level Linear Model to normalize Affymetrix HG-U95Av2 expression data from Gene Expression Omnibus for 16 disease-free prostate tissue samples, and 20 prostate tumor samples and their adjacent normal prostate tissue samples, using affylmGUI in R. The expression data used is a subset of the dataset record GDS2545 available at Gene Expression Omnibus[5]. GDS2545 is a dataset of normal prostate, prostate cancer, adjacent normal, and metastatic prostate cancer tissues analyzed on the HG-U95Av2 array from Affymetrix. affylmGUI is an open source software available through Bioconductor (http://www.bioconductor.org/index.html) that normalizes Affymetrix expression data using limma, another program available through bioconductor. Of the 94 genes associated with prostate cancer, 27 genes were annotated on the HG-U95Av2 chip. These 27 genes were used in the subsequent Gene Set Enrichment Analysis using the Java version at http://www.broadinstitute.org/gsea/. We also performed simulations using random sets of genes taken from the Database of Genomic Variants (http://projects.tcag.ca/variation) in order to account for potential Type 1 error due to possible association of hypervariable regions that may preferentially undergo genomic alterations in cancer. Using GSEA it was determined that this CNVR gene set showed enrichment in tumor tissue and in adjacent normal tissue compared to disease free normal tissue. This work was presented at the 59th Annual meeting of the American Society of Human Genetics, Honolulu HI. The genes that contribute to the core enrichment in both GSEA comparisons are shown in the column labeled “Overlap” in table 2. We have completed analyses of the identified CNVs in all of these genes except PSPC1 which we plan to do in the coming project year. In addition, 2 additional genes were enriched in the dataset comparing normal tissue to the tumor adjacent normal tissue. We will include these in our follow up.

We are currently seeking other expression datasets to perform similar analyses in hopes of prioritizing genes in these regions further. We are also conducting literature searches for

comparison of our gene set to those implicated in prostate cancer risk or progression in current studies.

Table 2. Results of GSEA with Genes in CNV regions.

Gene Set for GSEA		Leading Edge 'cancer vs. normal'	Leading Edge 'adjacent vs. normal'	Overlap	# Markers on Illumina array
UGT8	ARIH2	UGT8	UGT8	UGT8	3
HINT1	CYFIP1	HINT1	HINT1	HINT1	4
PTPRK	UQCRB	PTPRK	PTPRK	PTPRK	1
IGHM	TYRP1	IGHM	IGHM	IGHM	12
PPEF2	SETBP1	PPEF2	PPEF2	PPEF2	7
PTPRN2	TIPRL	PTPRN2	PTPRN2	PTPRN2	1
PSPC1	ADAM3A	PSPC1	PSPC1	PSPC1	4
AUTS2	ULK1	TYRP1	ULK1		
OTUD4	HINT1		HINT1		
COX10	BTG1		BTG1		
PPIE	ZBTB20		ZBTB20		
SLITRK3	GRIA1		AUTS2		
PDE1C	MAGI1				
LY6E					



	Normal vs. Adjacent	Normal vs. Tumor
Normalized Enrichment Score	-1.372	-1.497
Nominal p-value	0.091	0.036

KEY RESEARCH ACCOMPLISHMENTS

- We have developed independent assays to analyze and validate CNVs identified in prior work conducted in collaboration with deCode Genetics. Eight out of 8 assays conducted were validated.
- Using gene-set enrichment analysis, we have observed that the CNV regions identified using the array-based method are enriched for genes that have been observed to be differentially expressed in prostate tumor and normal tissue

REPORTABLE OUTCOMES

Blackburn A., Gelfond J., Goring H.H., Beuten Y., Thompson I., Leach RJ, and Lehman DM. (2009) Identification of Copy Number Variable Regions (CNVRs) Associated with Risk of Prostate Cancer in Mexican-Americans. Abstract presented at 59th Annual meeting of the American Society of Human Genetics, Honolulu HI, October 2009

CONCLUSION

Using gene set enrichment analysis, we have validated that copy number variant regions that differ between the Mexican American PCa cases and controls in SABOR are enriched for genes that are differentially expressed in prostate tumor tissue and in adjacent normal tissue as compared to disease free normal tissue. This supports our hypothesis that heritable structural variation may affect risk for PCa and/or its progression. We have confirmed the presence of 8 of these variants out of 8 tested using an independent assay. Three of these were not common in this population. The involvement of multiple rare variants in complex disorders has become widely accepted. Therefore, a comparison of the global burden of rare CNVs may help to elucidate effects. An analysis of biological pathways related to rare CNVs identified in this population may also uncover functional gene sets in PCa and further our knowledge of interactions between pathways that lead to cancer. This approach has recently proven successful in a study of autism spectrum disorders[4]. As genes are identified from these studies, they may prove to be both useful biomarkers for early diagnosis and/or excellent therapeutic targets for both prevention and treatment of prostate cancer.

Over this first project period we have gained an immense amount of experience in the complex area of genome-wide copy number variant identification and analyses. Although there is still no complete consensus on statistical analytical methods for determining copy number calls from SNP arrays, much progress has been made and 2 software programs have become the predominant choice. These are PennCNV[6] and QuantiSNP[7]. In another ongoing study of diabetes, we have applied these programs and their various tools to data from much denser Illumina arrays (500K – 1M duo arrays) for approximately 1200 Mexican American subjects from a local family-based cohort. We have been able to clearly distinguish breakpoints of CNVs due to the dense spacing of markers on these arrays. This has enabled better design of qPCR assays for confirmatory analyses. It also provides much greater confidence in copy calls

allowing the detection of rare variants as well as those that are small. Given our current experience with very dense SNP arrays, we see a need for applying these methods to this prostate cancer project. A comparison of 100 additional cases with 100 "hyper-normal" controls (i.e, older age) that are better matched for admixture using Illumina's OmniExpress array, which consists of approximately 750,000 probes, and the new statistical methodology will increase our ability to identify and better define copy variable regions, and importantly, those that are rare and/or small. We can then follow up statistically associated regions in fewer samples to validate. We anticipate a better assessment of copy variants to test for global burden as well as pathway analysis.

REFERENCES

1. Rose AE, Satagopan JM, Oddoux C et al. Copy number and gene expression differences between African American and Caucasian American prostate cancer. *J Transl Med* 2010;8(1):70.

2. Liedauer R, Svoboda M, Wlcek K et al. Different expression patterns of organic anion transporting polypeptides in osteosarcomas, bone metastases and aneurysmal bone cysts. *Oncol Rep* 2009 December;22(6):1485-92.

3. Kutuzov MA, Bennett N, Andreeva AV. Protein phosphatase with EF-hand domains 2 (PPEF2) is a potent negative regulator of apoptosis signal regulating kinase-1 (ASK1). *Int J Biochem Cell Biol* 2010 July 29.

4. Pinto D, Pagnamenta AT, Klei L et al. Functional impact of global rare copy number variation in autism spectrum disorders. *Nature* 2010 July 15;466(7304):368-72.

5. Chandran UR, Ma C, Dhir R et al. Gene expression profiles of prostate cancer reveal involvement of multiple molecular pathways in the metastatic process. *BMC Cancer* 2007;7:64.

6. Wang K, Li M, Hadley D et al. PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res* 2007 November;17(11):1665-74.

7. Colella S, Yau C, Taylor JM et al. QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. *Nucleic Acids Res* 2007;35(6):2013-25.

BIBLIOGRAPHY

Blackburn A., Gelfond J., Goring H.H., Beuten Y., Thompson I., Leach RJ, and Lehman DM. (2009) Identification of Copy Number Variable Regions (CNVRs) Associated with Risk of Prostate Cancer in Mexican-Americans. Abstract presented at 59th Annual meeting of the American Society of Human Genetics, Honolulu HI, October 2009

Personnel: Donna Lehman, Robin Leach, Jon Gelfond, Christopher Loudon, August Blackburn